TELECOM CHURN

Problem Statement

- In the telecom industry, customers are able to choose from multiple service providers and actively switch from one operator to another. In this highly competitive market, the telecommunications industry experiences an average of 15-25% annual churn rate. Given the fact that it costs 5-10 times more to acquire a new customer than to retain an existing one, customer retention has now become even more important than customer acquisition.

- For many incumbent operators, retaining high profitable customers is the number one business goal.

- To reduce customer churn, telecom companies need to predict which customers are at high risk of churn.

- In this project, you will analyze customer-level data of a leading telecom firm, build predictive models to identify customers at high risk of churn and identify the main indicators of churn.

STEPS OF THE PROJECT

- Data Reading
- Exploratory Data Analysis and Data Cleaning
- Data Visualization
- Feature Importance
- Feature Engineering
- Setting a baseline
- Splitting the data in training and testing sets
- Assessing multiple algorithms
- Algorithm selected: Gradient Boosting
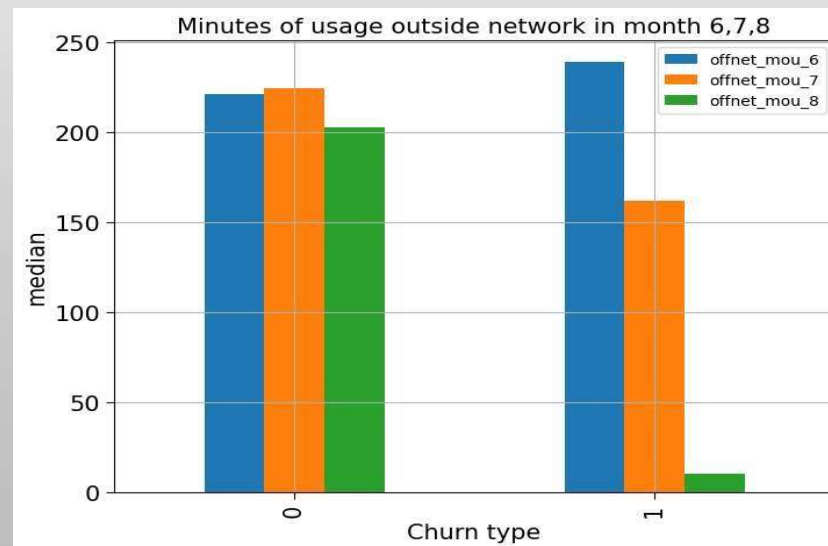- Hyper parameter tuning
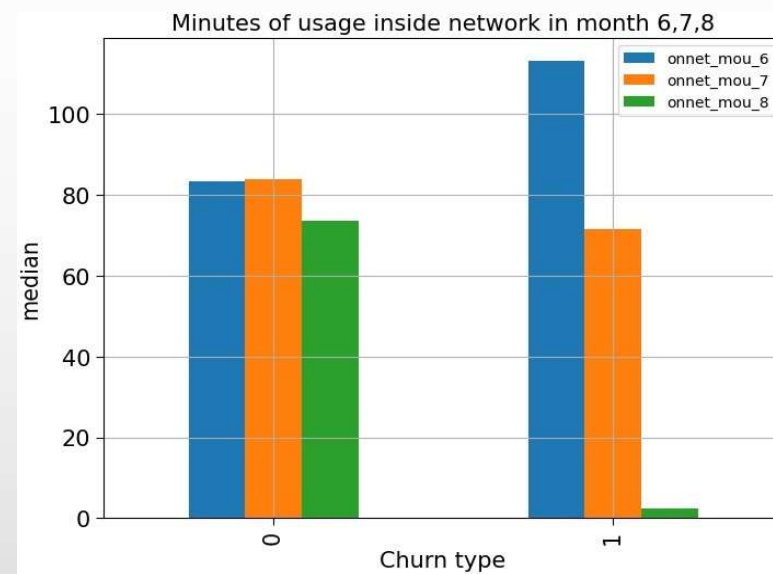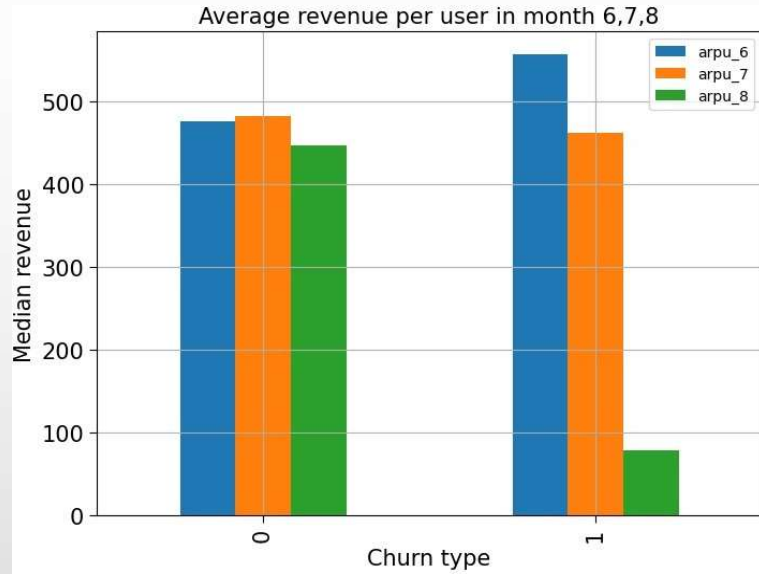- Performance of the model
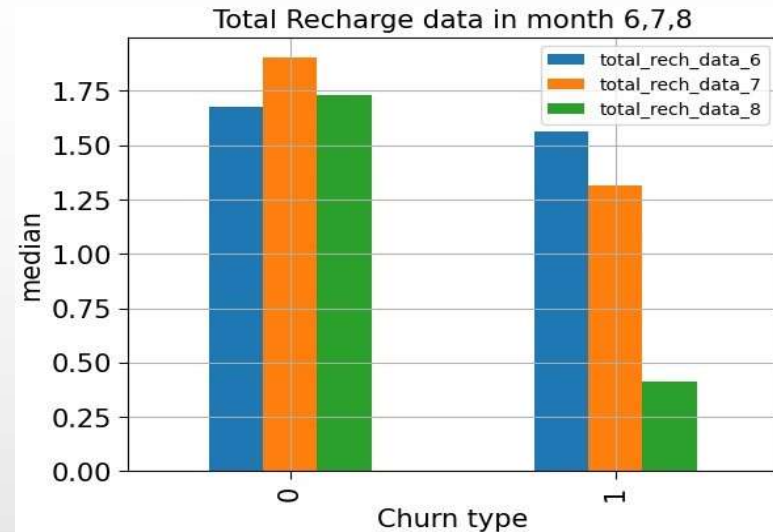- Drawing conclusions - Summary
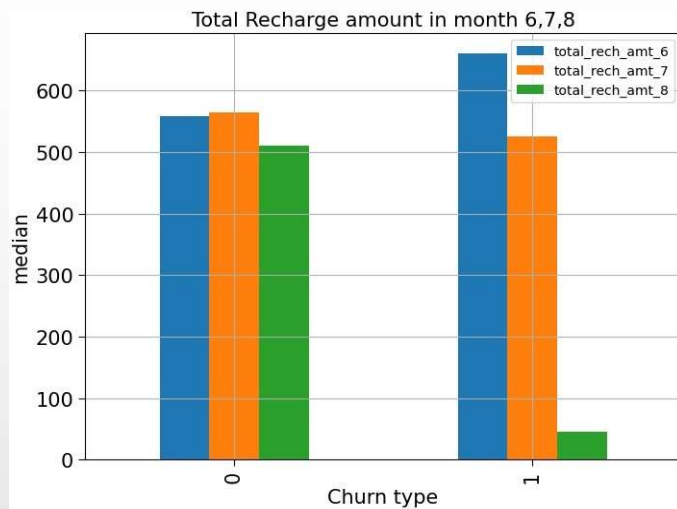
Data Reading And Cleaning

- Dataset contains 99999 no of rows.
- 226 no of columns.
- Number of Float data type - 179
- Number of int data type – 35
- Number of object data type- 12

EXPLORATORY DATA ANALYSIS

Churn and Non-Churn distributions in percentage
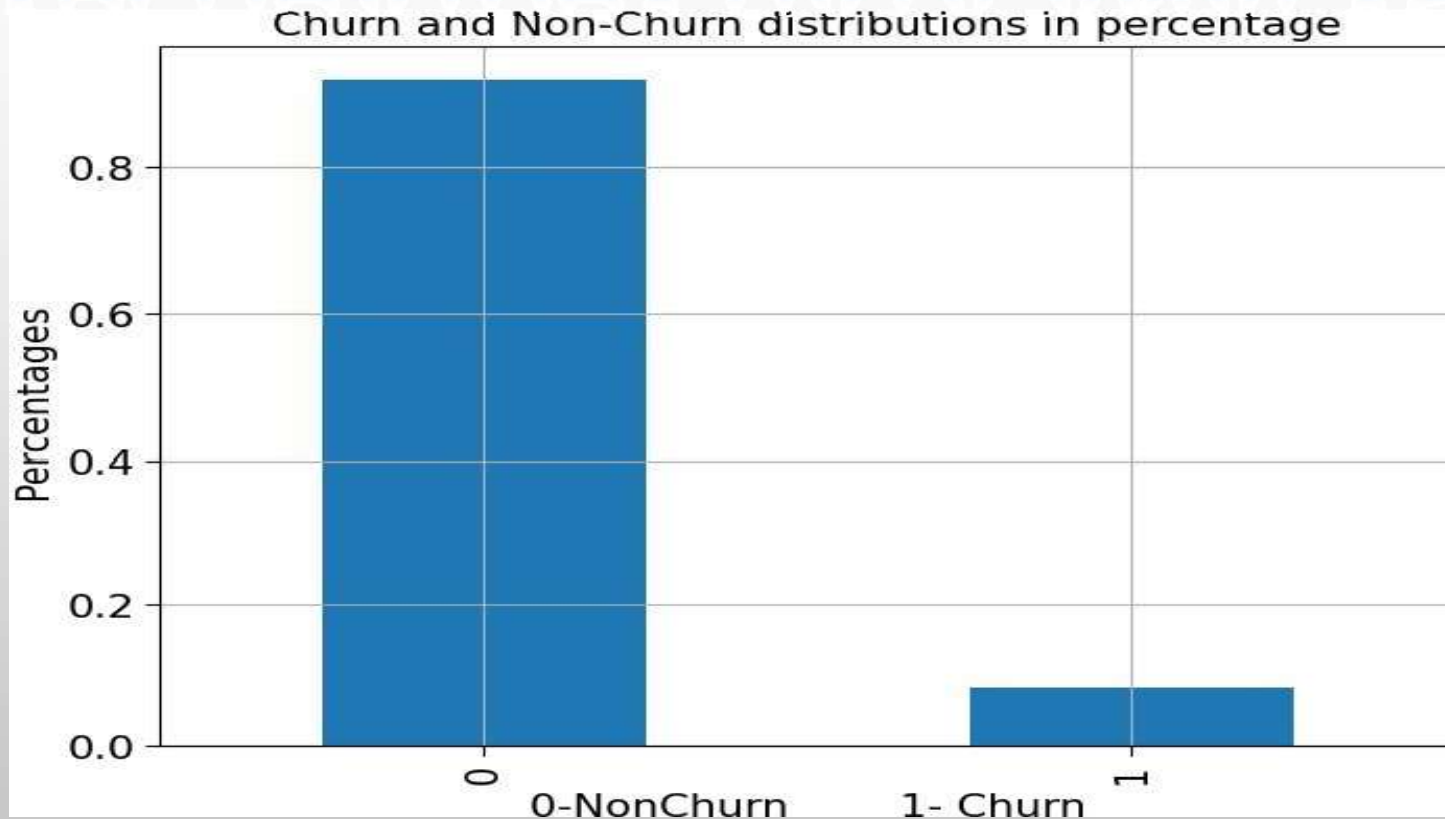
•We have 92% customers belong non–churn and 8% customers belong to Churn type.

Average revenue per user in month 6,7,8

Minutes of usage inside network in month 6,7,8

Minutes of usage outside network in month 6,7,8

Total Recharge amount in month 6,7,8

Total Recharge data in month 6,7,8

•Average revenue per user more in month 6 means, if they are unsatisfied, those users are more likely to churn.

•Users whose minutes of usage are more in month 6, they are more likely to churn.

•The users who have big difference of minutes of call duration to other network between month 6 and month 7,are likely to churn.

•when the difference of total recharge amount is more, those users are more likely to churn.

•Users who have not recharge in month 6, 7, 8 may or may not churn, we do not have much evidence from data.
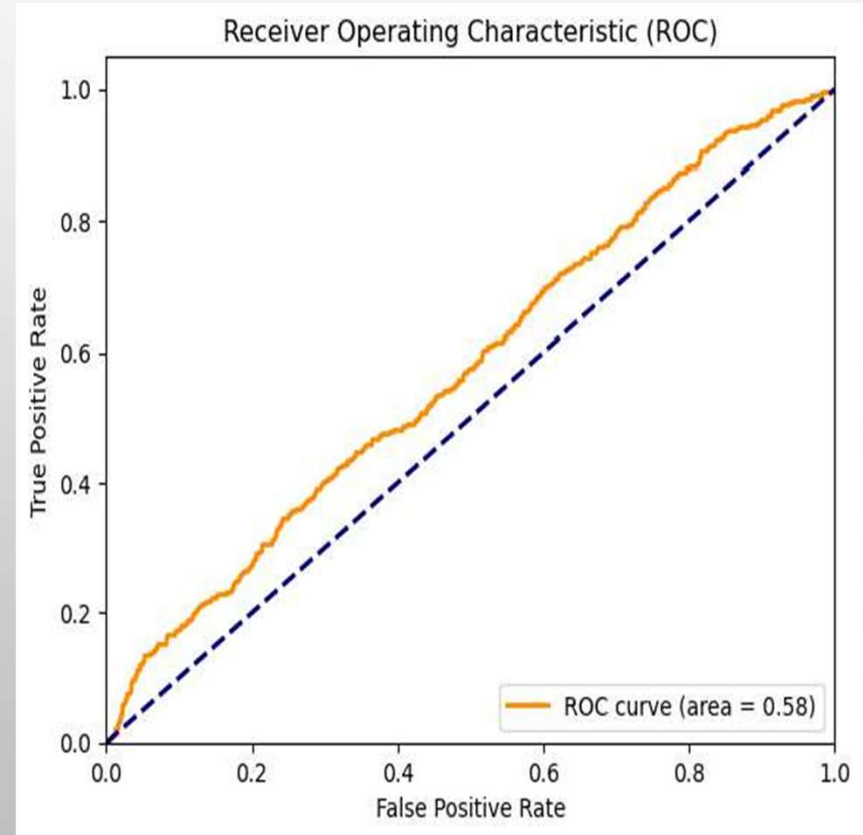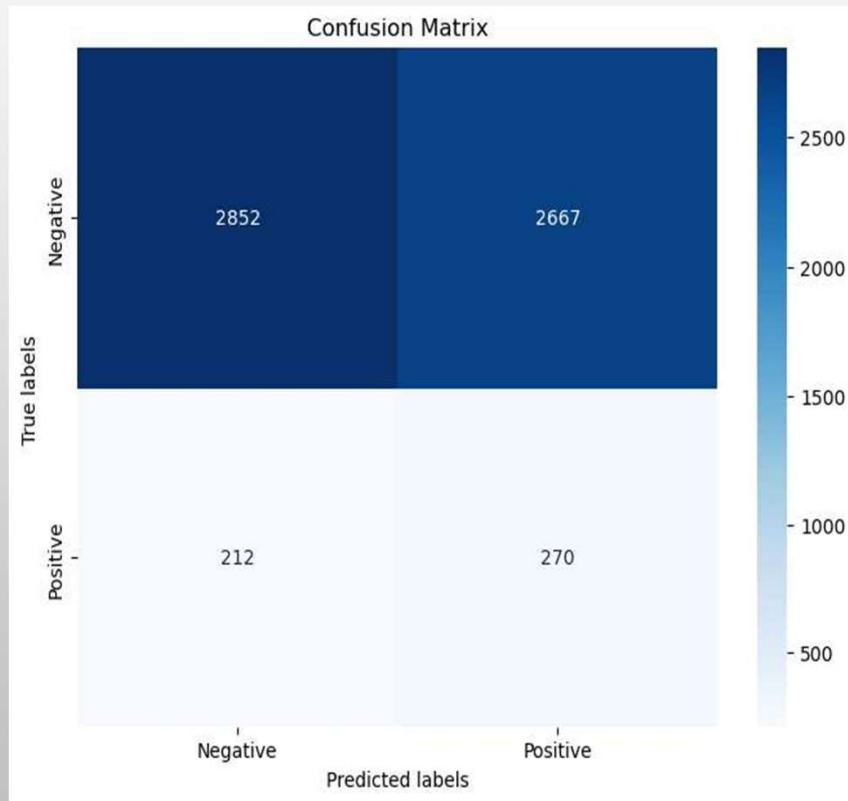
MODEL BUILDING:
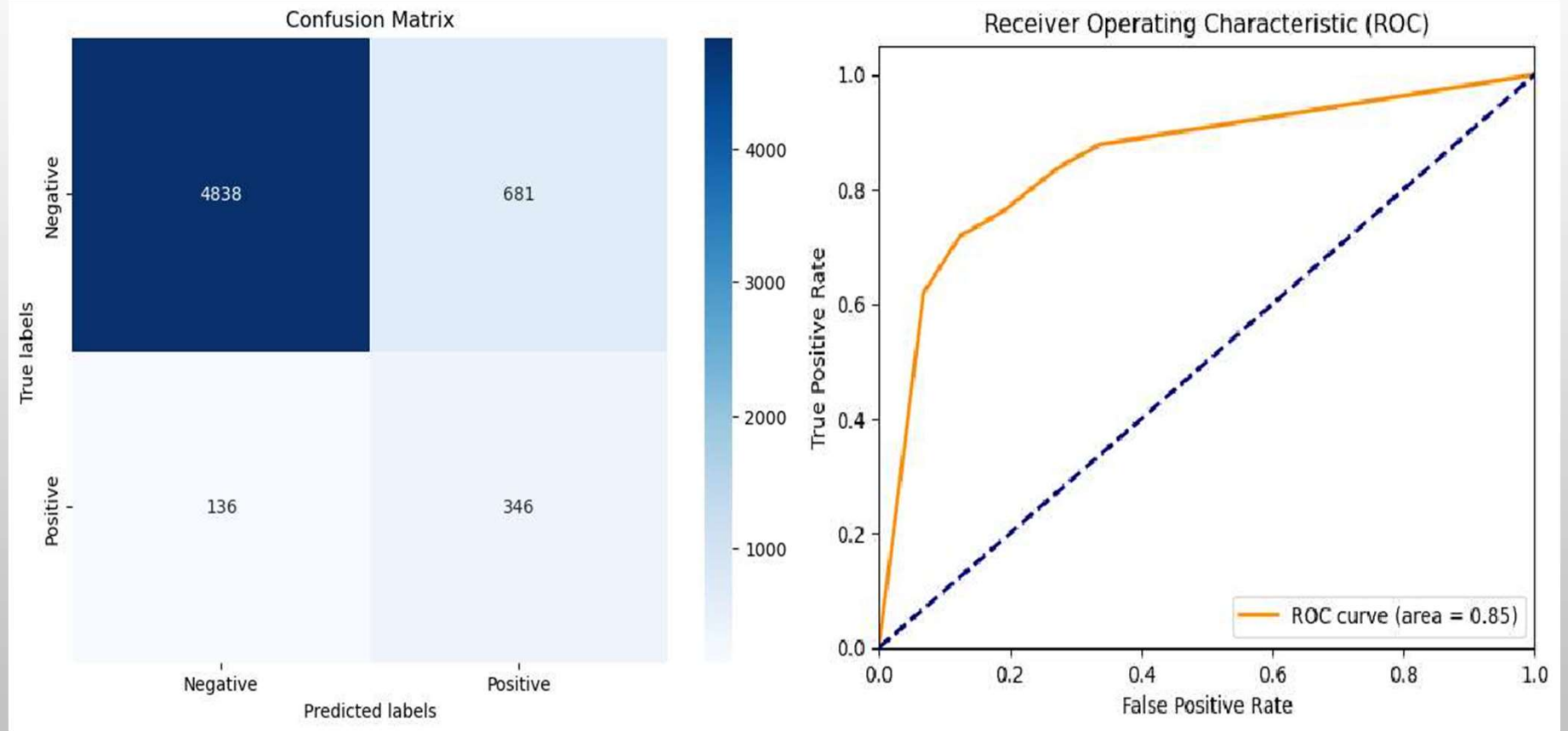
We will explore below models.
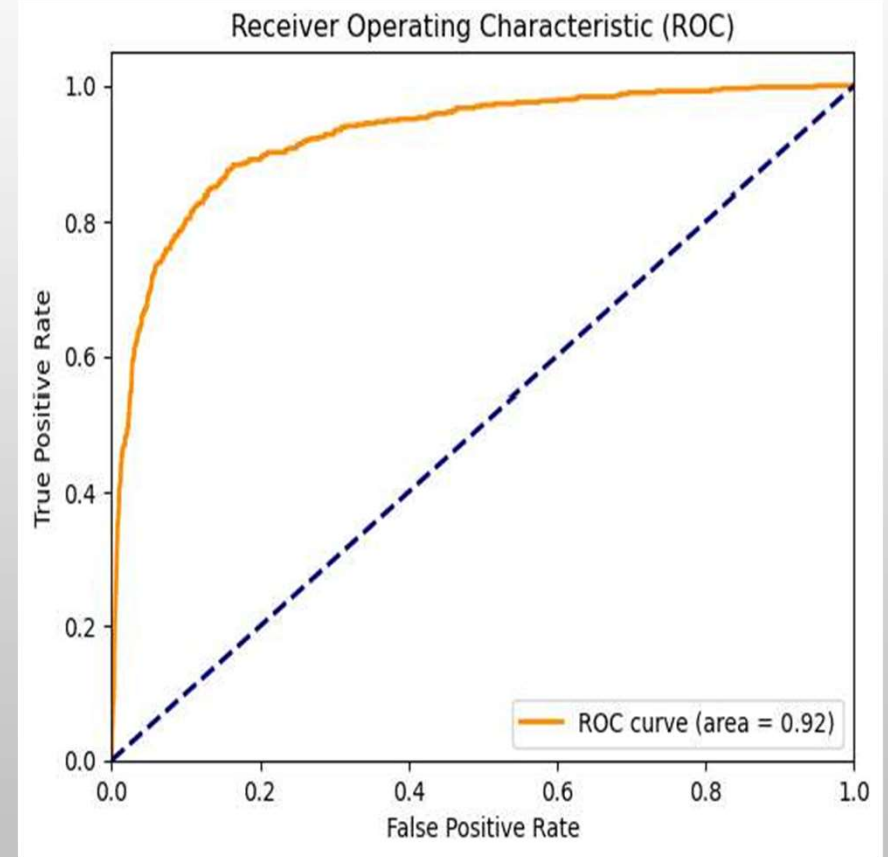
- Logistic regression
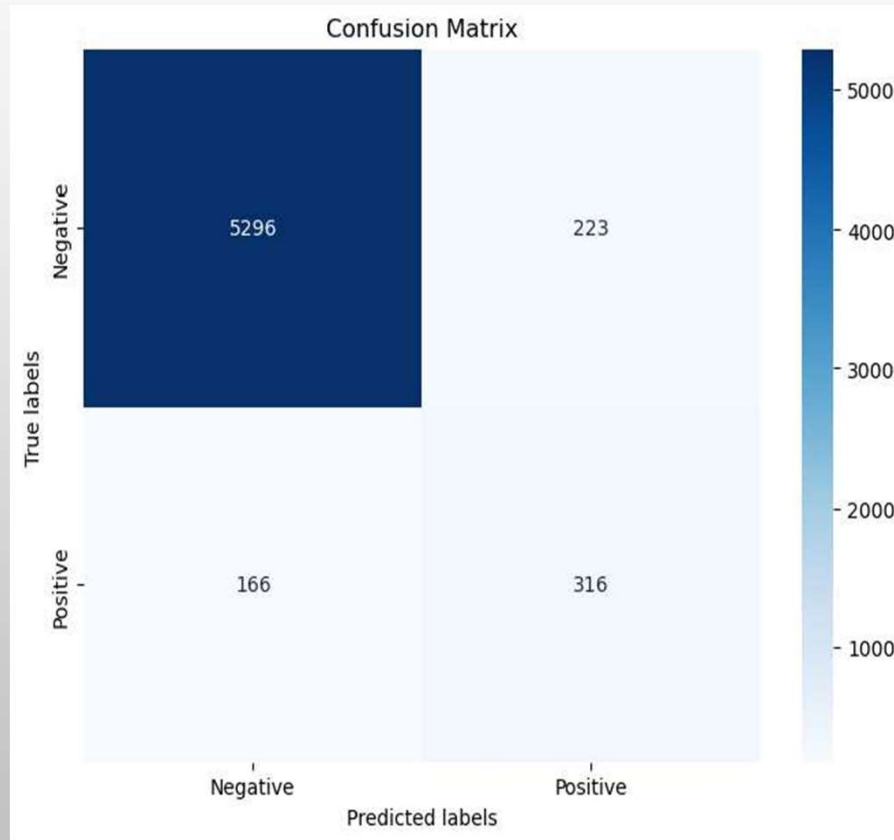- Decision tree
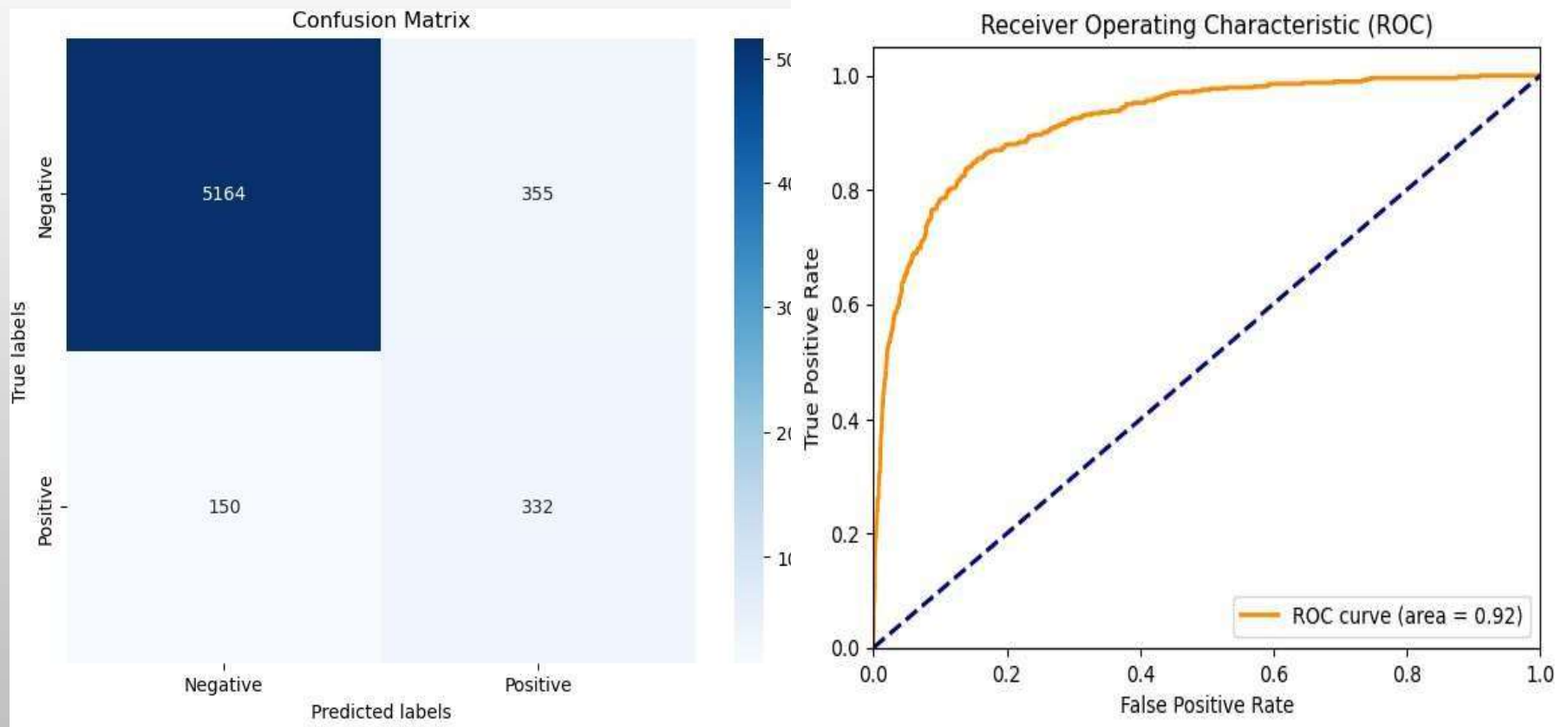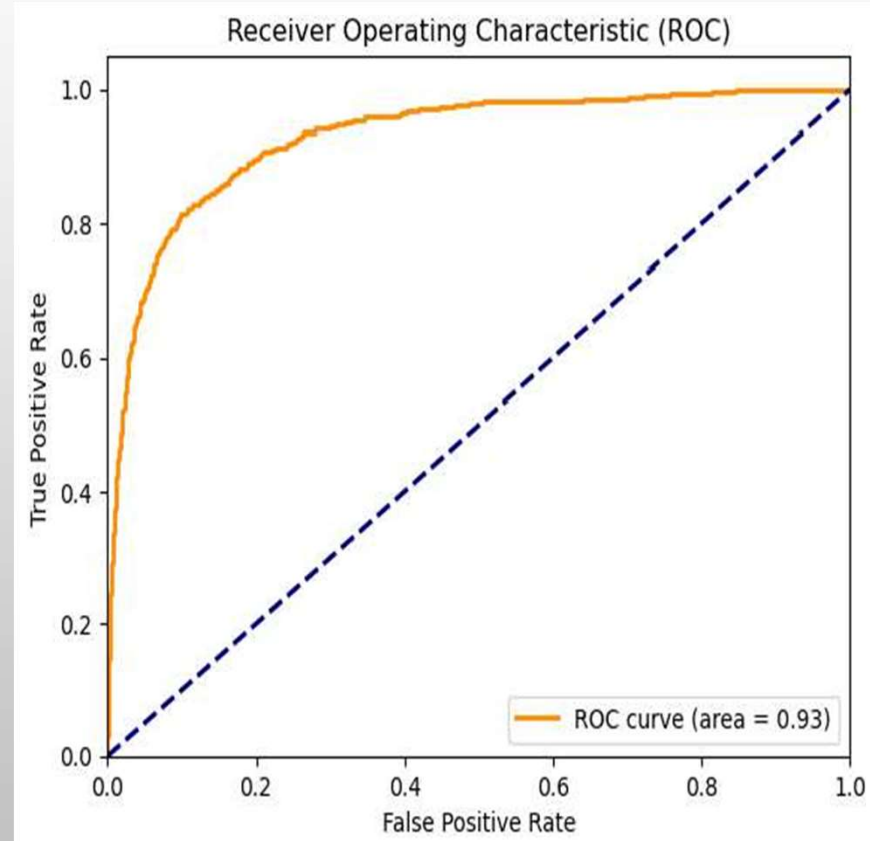- Randomforest
- Gradientboosting
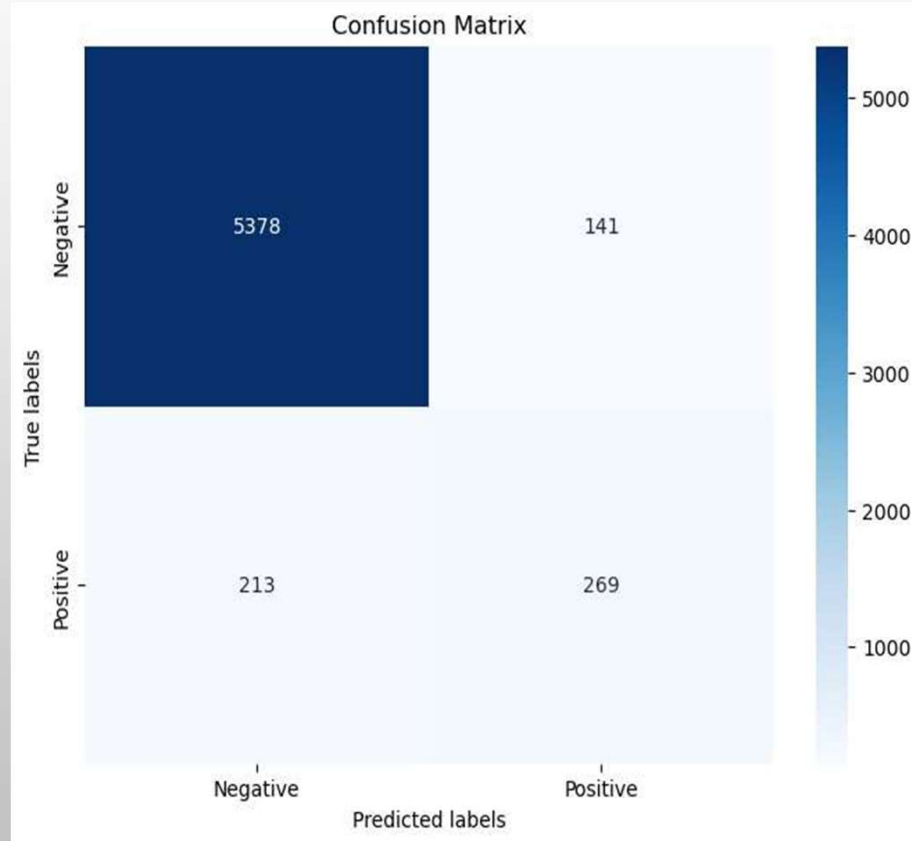- XGboost

# LOGISTIC REGRESSION

DECISION TREE

RANDOMFOREST

GRADIENTBOOSTING

XGBOOST

	Model	precision	recall	f1_score	roc_auc
0	LogisticRegression	0.091931	0.560166	0.157941	0.575711
0	DecisionTree	0.336904	0.717842	0.458582	0.851062
0	RandomForest	0.586271	0.655602	0.619001	0.924402
0	GradientBoosting	0.483261	0.688797	0.568007	0.919795
0	XGBoost	0.656098	0.558091	0.603139	0.929580

•The random forest worked well on this data in churn with precision close to 59%, recall close to 65% and f1_score close to 61%.
•In Logistic regression we have used PCA.
•In this scenario, Without PCA model works well.

FEATURE IMPORTANCE AND MODEL INTERPRETATION

Feature Importance
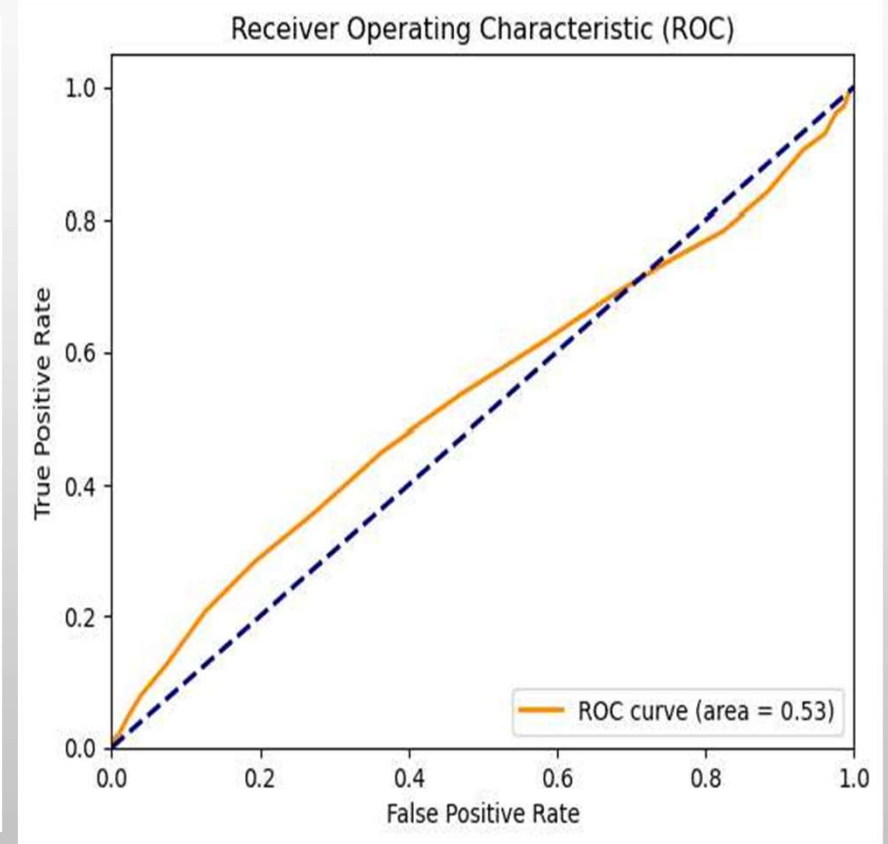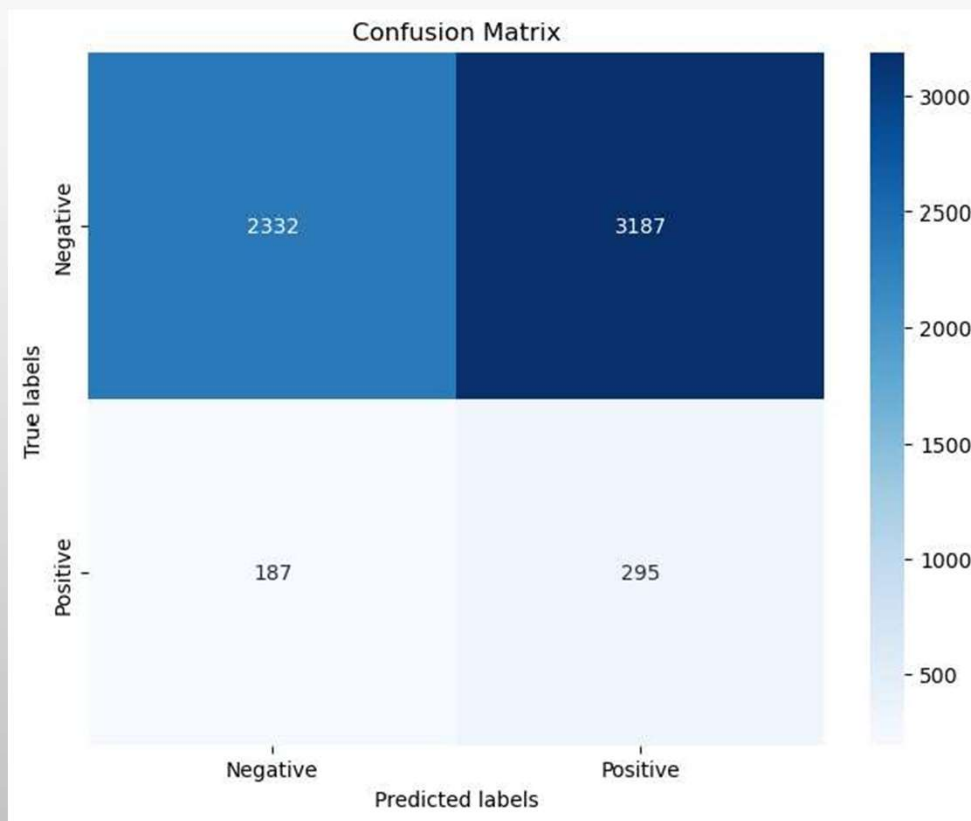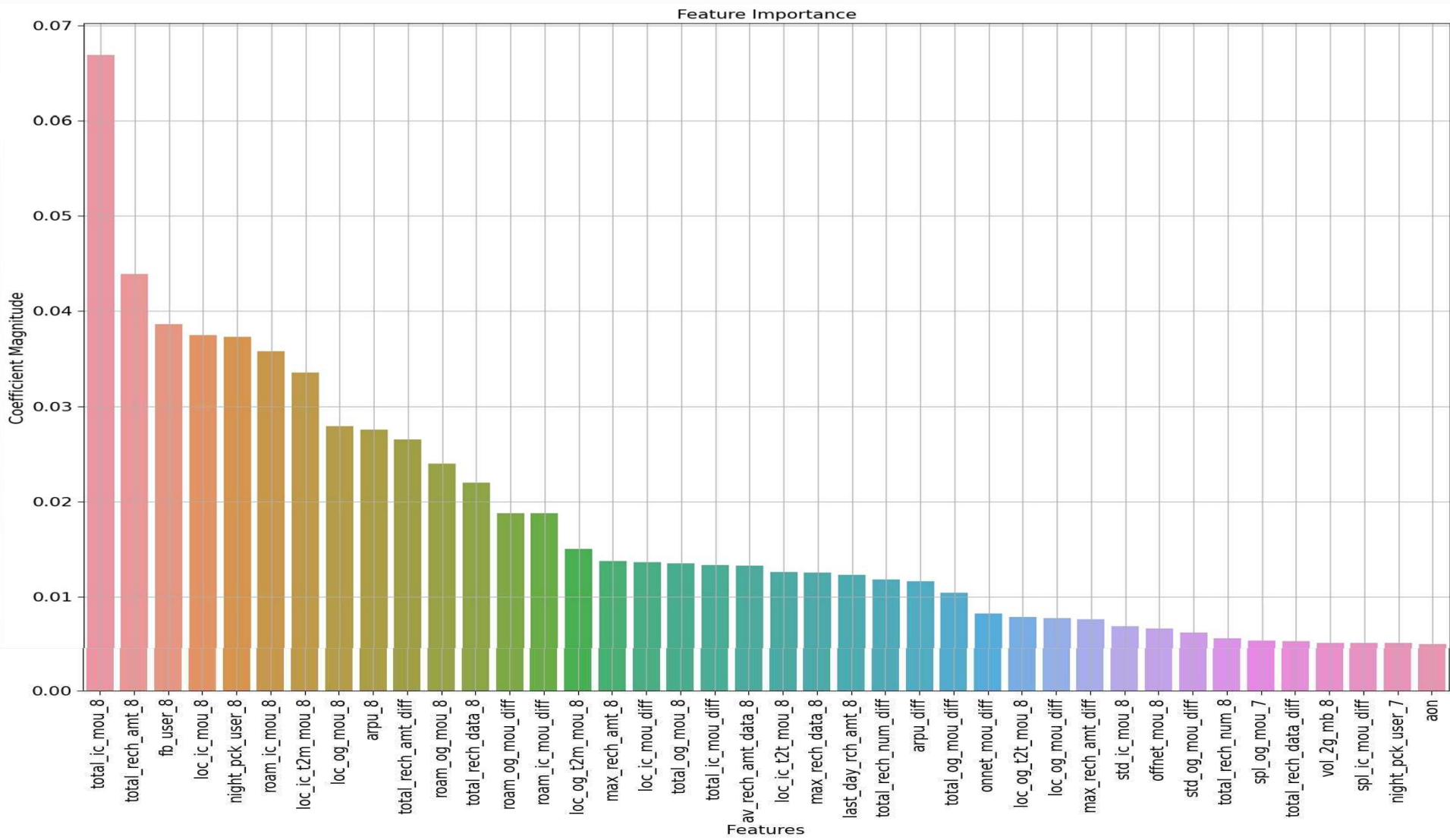
CONCLUSIONS:

•The most important features are as shown in above graph.

•Average revenue per user more, those are likely to churn if they are not happy with the network.

•local calls minutes of usage has also has impact on churn .

•Large difference between recharge amount between 6th and 7th month, also impact churn.

•Users who are using more Roaming in Outgoing and Incoming calls, are likely to churn. Company can focus on them too.